WHAT IS CLAIMED IS:

1. A method for throttling incoming requests to a server having a buffer for storing incoming calls request to being processed by the server, the method comprising:

receiving incoming requests;

determining whether the server is overloaded;

if the server is overloaded then storing the incoming requests in the buffer only if a number of incoming requests currently stored in the buffer is less than an acceptance limit; and

if the server is not overloaded, then storing the incoming request in the buffer.

2. The method according to claim 1, wherein determining whether the server is overloaded is based on at least an incoming request rate and a processing rate of the server.

3. The method according to claim 2, wherein the server is overloaded if the incoming request rate exceeds the processing rate of the server.

4. The method according to claim 3, wherein the server is overloaded if the incoming request rate exceeds the processing rate of the server for a predetermined amount of time.

5. The method according to claim 4, wherein the predetermined amount of time is set based on a desired delay time of the buffer.

6. The method according to claim 1, wherein the acceptance limit is less than 100% of a capacity of the buffer while the server is overloaded.

7. The method according to claim 1, wherein the acceptance limit is 25% of the capacity of the buffer.

8. The method according to claim 1, wherein the incoming requests are telephone calls.

9. The method according to claim 1, wherein the incoming requests are data transmissions.

10. The method according to claim 1, wherein the acceptance limit is determined based on a desired delay time of the buffer.

11. An apparatus for throttling incoming requests, comprising:

a server having a buffer that stores incoming calls request to being processed by the server; and

a controller coupled to the server that determines whether the server is overloaded,

if the server is overloaded then storing the incoming requests in the buffer only if a number of incoming requests currently stored in the buffer is less than an acceptance limit, and

if the server is not overloaded, then storing the incoming request in the buffer.

12. The apparatus according to claim 11, wherein determining whether the server is overloaded is based on at least an incoming request rate and a processing rate of the server.

13. The apparatus according to claim 12, wherein the server is overloaded if the incoming request rate exceeds the processing rate of the server.

14. The apparatus according to claim 13, wherein the server is overloaded if the incoming request rate exceeds the processing rate of the server for a predetermined amount of time.

15. The apparatus according to claim 14, wherein the predetermined amount of time is set based on a desired delay time of the buffer.

16. The apparatus according to claim 11, wherein the acceptance limit is less than 100% of a capacity of the buffer.

17. The apparatus according to claim 11, wherein the acceptance limit is 25% of the capacity of the buffer.

18. The apparatus according to claim 11, wherein the incoming requests are telephone calls.

19. The apparatus according to claim 11, wherein the incoming requests are data transmissions.

20. The apparatus according to claim 11, wherein the acceptance limit is determined based on a desired delay time of the buffer.